1-1-2020

Missing data imputation of high-resolution temporal climate time series data

Eben Afrifa-Yamoah
Edith Cowan University, e.afrifayamoah@ecu.edu.au

Ute A. Mueller
Edith Cowan University, u.mueller@ecu.edu.au

S. M. Taylor

A. J. Fisher
Edith Cowan University

10.1002/met.1873
Afrifa-Yamoah, E., Mueller, U. A., Taylor, S. M., & Fisher, A. J. (2020). Missing data imputation of high-resolution temporal climate time series data. Meteorological Applications, 27(1), e1873. https://doi.org/10.1002/met.1873

Received: 19 November 2018 | Revised: 29 September 2019 | Accepted: 2 December 2019 | Published on: 9 January 2020

DOI: 10.1002/met.1873

RESEARCH ARTICLE

Meteorological Applications RMetS

Missing data imputation of high-resolution temporal climate time series data

E Afrifa-Yamoah[1] | **U. A. Mueller**[1] | **S. M. Taylor**[2] | **A. J. Fisher**[1]

[1]School of Science, Edith Cowan University, Joondalup, Australia

[2]Department of Primary Industries and Regional Development (DPIRD), Western Australian Fisheries and Marine Research Laboratories, North Beach, Australia

Correspondence
Ebenezer Afrifa-Yamoah, School of Science, Edith Cowan University, 270 Joondalup Drive, Joondalup, WA 6027, Australia.
Email: e.afrifayamoah@ecu.edu.au

Funding information
WA Department of Primary Industries and Regional Development

Abstract

Analysis of high-resolution data offers greater opportunity to understand the nature of data variability, behaviours, trends and to detect small changes. Climate studies often require complete time series data which, in the presence of missing data, means imputation must be undertaken. Research on the imputation of high-resolution temporal climate time series data is still at an early phase. In this study, multiple approaches to the imputation of missing values were evaluated, including a structural time series model with Kalman smoothing, an autoregressive integrated moving average (ARIMA) model with Kalman smoothing and multiple linear regression. The methods were applied to complete subsets of data from 12 month time series of hourly temperature, humidity and wind speed data from four locations along the coast of Western Australia. Assuming that observations were missing at random, artificial gaps of missing observations were studied using a five-fold cross-validation methodology with the proportion of missing data set to 10%. The techniques were compared using the pooled mean absolute error, root mean square error and symmetric mean absolute percentage error. The multiple linear regression model was generally the best model based on the pooled performance indicators, followed by the ARIMA with Kalman smoothing. However, the low error values obtained from each of the approaches suggested that the models competed closely and imputed highly plausible values. To some extent, the performance of the models varied among locations. It can be concluded that the modelling approaches studied have demonstrated suitability in imputing missing data in hourly temperature, humidity and wind speed data and are therefore recommended for application in other fields where high-resolution data with missing values are common.

KEYWORDS

high-resolution climate time series data, imputation, missing observations, short cycle duration, state-space modelling

1 | INTRODUCTION

Climatic conditions such as precipitation, temperature, humidity, wind speed, wind gust and sea level pressure have been used over time in many meteorological, energy application, agricultural, ecological and hydrological studies (Firat *et al*., 2012; Xu *et al*., 2013; Lara-Estrada *et al*., 2018). Weather stations across the world continue to record and monitor various climatic parameters for climate classification, planning, modelling and management purposes (Firat *et al*., 2012; Lara-Estrada *et al*., 2018). In recent years, the threats of global warming and climate change (World Bank, 2012) have sparked a resurgent interest in the analysis and inference of climatic variables and related subjects in the natural, social and political sciences. Climate data are typically processed and analysed at low-resolution levels such as daily, weekly, monthly and yearly resolution (Firat *et al*., 2012, Kanda *et al*., 2018). In contrast, processing and analysing data at a high-resolution scale such as h minutes ($h \leq 60$) results in the availability of an appreciable number of points even when the overall time period under investigation is short; e.g. a (leap) year-long hourly time series consists of 8,784 data points. Moreover, the analysis of high-resolution data offers greater ability to understand the nature of data variability, behaviours, trends and to detect small changes (Pincetl *et al*., 2015). In many instances, high-resolution climatic data are incomplete. Yozgatligil *et al*. (2013) attributed the gaps to faulty measuring instruments, which are caused by recording errors or malfunctioning equipment, or instances of meteorological extremities and remoteness, routine maintenance and sensor calibration (Coble *et al*., 2012). In effect, missing observations in climate data often occur consecutively for long periods of time (Simolo *et al*., 2010).

Climate studies require complete time series data which, in the presence of missing data, means that imputation must be undertaken. A literature search identified that several imputation methods have been applied to relatively low-resolution climate data, typically of daily, monthly and yearly resolutions. The unconditional mean imputation is very simple in application; however, results suggest that it is not robust and often results in an underestimation of the standard errors (Yozgatligil *et al*., 2013). The expectation maximization (EM) algorithm developed by Dempster *et al*. (1977) has been applied to extrapolate missing data in an analysis of monthly temperature time series, e.g. by Firat *et al*. (2012). Simolo *et al*. (2010) proposed the completion method with the ability to preserve the probability density function in imputing missing values in a daily precipitation dataset. Xu *et al*. (2013) applied the point estimation model of Biased Sentinel Hospitals-based Area Disease Estimation (P-BSHADE) to interpolate missing data in an annual temperature dataset. In a comparative study on the estimation of missing values in daily temperature and precipitation data, Kanda *et al*. (2018) reported that multiple regression using the least absolute deviation performed best based on four performance indicators, namely mean absolute error (MAE), root mean squared error (RMSE), coefficient of efficiency and skill score. Other methods such as the regularized EM algorithm (Schneider 2001), the Fourier fit, the artificial neural network (McCandless *et al*. 2011; Kashani and Dinpasho, 2012), the multilayer perceptron neural network, the EM-Markov chain Monte Carlo (Yozgatligil *et al*., 2013) and the Bayesian network (Lara-Estrada *et al*., 2018) have been applied in the imputation of missing observations in daily and monthly precipitation, temperature and humidity data.

The analysis of low-resolution climate data is commonly based on aggregation from higher-resolution datasets. It is important to note that, for incomplete datasets, the estimation of fundamental statistics such as the mean and covariance is challenging, mostly biased and can be misleading (Schneider, 2001). Therefore, in a field where missing observations are common, there should be clarity about how low-resolution data are derived and how missing values are handled. The effective handling of missing observations in finer-resolution data would guarantee a precise estimation of parameters to guide decision making processes. To the best of our knowledge, in the literature on the imputation of missing data in meteorological settings only low-resolution climate datasets were considered. Although there is some literature on imputation for high-resolution data in other areas, such as electricity consumption (Hyndman and Fan, 2015), audio signal processes (Smaragdis *et al*., 2009) and electrochemical ozone measurements (Pang *et al*., 2017), the underlying mechanisms of their data generating processes are different from those of climate datasets. The gap that needs to be addressed includes investigation and imputation of missing data in high-resolution climate datasets. This study evaluated the performance of univariate time series models by state-space methods and multiple linear regression models driven by other climatic variables to impute missing values in a 12 month time series of hourly measurements from four locations in Western Australia (WA). The climatic variables considered were temperature, humidity and wind speed.

In dealing with the imputation of missing data, the fundamental principles are to understand and use the nature of the data including the cause for the missing data occurrences. Climatic data are generally characterized by properties such as autocorrelation between time lags, seasonality, periodic trends, cycles and the homogeneity effect over geographical areas. These characteristics

are profound in high-resolution data and could be useful information for developing imputation modelling schemes to "fill in" the periods of missingness (Moritz *et al*., 2015). According to Rubin (1976), missingness in data may be completely at random (MCAR, data are missing independently of both observed and unobserved data), random (MAR, given the observed data, data are missing independently of unobserved data) or non-random (MNAR, missing observations are related to values of unobserved data). There are also cases where the missing mechanism is censored (De Jong *et al*., 2016). Most imputation methods assume MCAR and MAR because their missing data mechanisms are said to be ignorable, implying that the imputation modelling does not require special models for the cause of missingness and the likely values for the gaps (Rubin, 1976; Moritz *et al*., 2015). However, a clearer understanding of the cause of missingness helps to infer the distribution of the data gaps in the dataset and would enable the design of reasonable simulators on test data for validation purposes (Moritz *et al*., 2015).

The development of MCAR and MAR imputation schemes relies on the nature of the relationship between variables. For instance, MAR driven models involving highly correlated variables typically yield highly plausible imputation values (van Buuren and Groothuis-Oudshoorn, 2011). In the case of univariate time series, MCAR and MAR imputations are similar and draw inferences on the behaviour of the variable over time (Moritz *et al*., 2015). In the modelling phase, there is a need to address the high levels of stochasticity in most climate time series. This can be achieved by applying models whose parameters evolve over time (Harvey, 1989). The classical autoregressive integrated moving average (ARIMA) and seasonal ARIMA models have been applied to the imputation of missing data in univariate time series (Yodah *et al*., 2013). However, structural time series models, also known as unobserved component models, have evolutionary properties to study dynamic phenomena including imputations (Moritz *et al*., 2015). Both additive and multiplicative time series can be modelled structurally. For instance, multiplicative models can be additively decomposed via the logarithm of model components and each component can be fitted with an independent autoregressive moving average or ARIMA process (Pollock, 2008). To capture the complexity of the behaviour of climate time series over time adequately and to fine-tune the parameter estimation process, smoothing functions with the ability to extract the signals in the time series can be jointly modelled with the state-space representation of time series models for prediction and imputation purposes (Moritz and Bartz-Beielstein, 2017). The common smoothing functions in time series applications are the Kalman filter and the locally weighted scatterplot smoother (lowess) curve (Bianchi *et al*., 1999; Moritz *et al*., 2015, Moritz and Bartz-Beielstein, 2017). The Kalman filter is often preferred because it uses maximum likelihood to estimate the smoothing parameters and tends to give optimal results since the mean square errors of one-step-ahead predictions are minimized and also for its ability to provide sequentially updated estimates (Bianchi *et al*., 1999).

Alternatively, regression-based approaches such as the ordinary, lagged, Bayesian multiple linear regression and dynamic regression models are well-known modelling techniques for prediction and in cases of highly correlated variables produce better imputation estimates (Petris, 2010; van Buuren and Groothuis-Oudshoorn, 2011). Climate variables are generally correlated, e.g. temperature and humidity, and, in instances where weather stations collect multiple climatic conditions, imputations for missing observations in one missing variable can be inferred from the other observed variables.

2 | METHODS

This section presents the study area, data description and simplified versions of the modelling techniques evaluated in the imputation. Advanced time series models with state-space representation amenable to Kalman filter and smoothing algorithms (Pollock, 2008) are discussed. More detailed information about the time series modelling techniques presented below can be found in Harvey (1989), Harvey and Peters (1990), Pollock (2008) and Jalles (2009).

2.1 | Study area and data description

WA has a coastline stretching 12,889 km and 10 different climatic zones according to the Köppen–Geiger climate classification (see Figure 1). The Australian Government Bureau of Meteorology collects weather observation data from stations in 14 districts across WA. In this study, climate data from four locations (Esperance [009789], Perth [009105], Learmonth [005007] [henceforth Exmouth] and Broome [003003]; Figure 1) were analysed between March 1, 2011 and February 29, 2012. These coastal weather stations were chosen because remote camera monitoring also occurs at these sites as part of ongoing research on recreational fishing (Steffe *et al*., 2017) and each site is located in one of the four marine bioregions off the WA coastline: Esperance in the South Coast region, Perth in the West Coast region, Exmouth in the Gascoyne Coast region and Broome in the North Coast (Ryan *et al*., 2015).

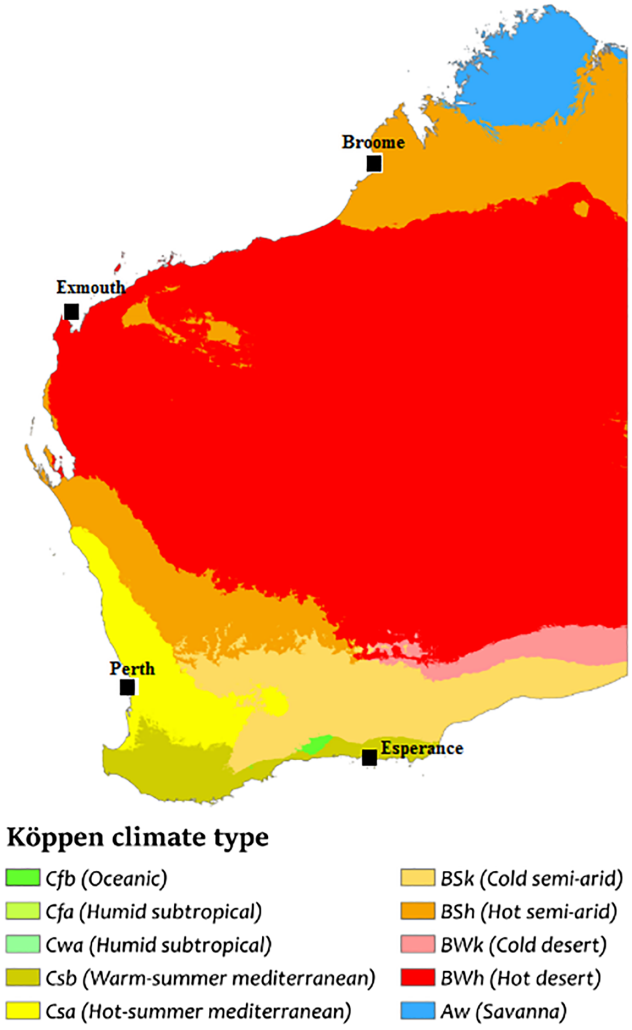

FIGURE 1 Western Australia's climate classification

Precipitation, temperature, humidity, wind speed and direction and sea level pressure were provided by the Australian Bureau of Meteorology at an hourly resolution. The overall time period covered was March 1, 2011 to February 29, 2012. The imputation focused on missing observations in the temperature, humidity and wind speed time series data while the other variables were used as predictors in the modelling phase. The climatic variables analysed in this study will be used as a principal source of inference in building models to impute missing data from the remote cameras that record recreational boating activity related variables.

The study methodology adopted for the purposes of this study was five-fold cross-validation. In this scheme, missing data were imputed for five different folds of missing patterns and the resulting imputations were compared to the true values. For each of the four locations the time interval of greatest duration without missing data in the original dataset was identified. These complete sub-samples of hourly temperature, humidity and wind speed data formed the basis of our analysis. The assumption that data were missing at random (MAR) was imposed and artificial gaps of missing observations were created in the complete sub-samples to mimic the general nature of missing patterns in the records provided by the Australian Bureau of Meteorology. The percentage of missing data was set to 10%, as in the datasets provided for the four locations none recorded more than 10% missing observations. Table 1 presents the dates and data points for the four sub-samples and the number of missing values (NA) created.

TABLE 1 Description of sub-samples and the number of missing values

Location	Date	Data points	NA
Esperance	27 Mar to 31 Oct 2011	5,242	524
Perth	17 Oct 2011 to 29 Feb 2012	3,984	398
Exmouth	16 Mar to 17 Oct 2011	4,447	445
Broome	01 Mar to 22 Aug 2011	4,200	420

The distribution of the lengths of gaps of missing observations in the five missing patterns is displayed in Figure 2.

2.2 | Autoregressive integrated moving averages (ARIMAs)

ARIMA models are perhaps the most used time series technique in an application. The models are theoretically and statistically sound, flexible and make few assumptions (Ho and Xie, 1998). For a univariate time series y_t, an ARIMA(p, d, q) is defined as:

$$\theta_p(B)(1-B)^d y_t = \phi_q(B)a_t \tag{1}$$

$$\theta_p(B) = 1 - \sum_{i=1}^{p} \theta_i B^i \quad \text{where } \theta_p \neq 0 \tag{2}$$

$$\phi_q(B) = 1 - \sum_{i=1}^{q} \phi_i B^i \quad \text{where } \phi_q \neq 0 \tag{3}$$

$$B^n y_t = y_{t-n} \tag{4}$$

where a_t is a white noise process, p and q denote the order of the autoregressive and moving average processes, B is the backshift operator and d is the order of differencing. The vector $\boldsymbol{\theta} = (\theta_1, ..., \theta_p)$ denotes the vector of p parameters for the autoregressive process and $\boldsymbol{\phi} = (\phi_1, ..., \phi_q)$ is the vector of q parameters for the moving average process. These explain the nature of the autocorrelation

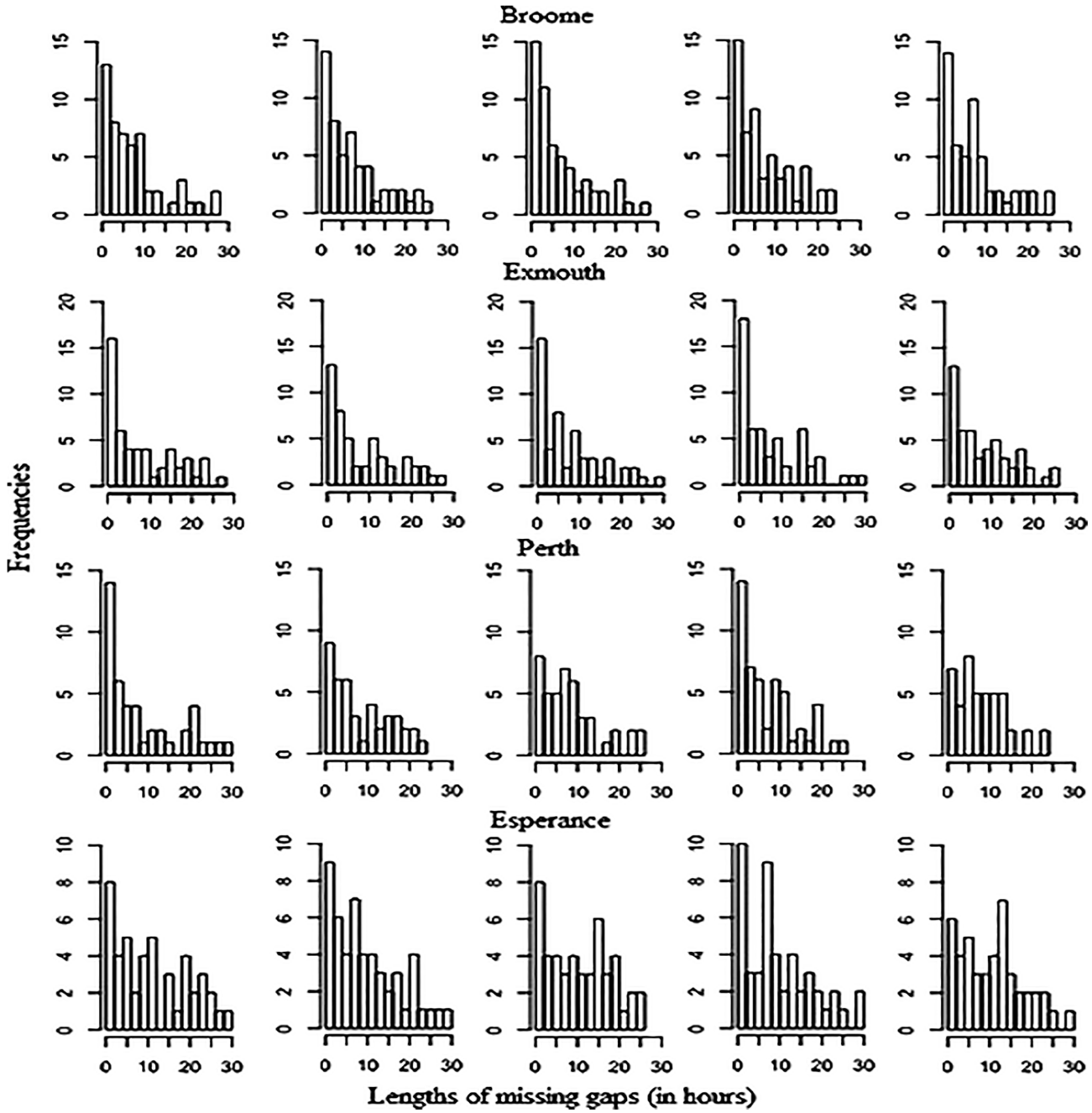

FIGURE 2 The distribution of the lengths of gaps measured in hours in the five folds in the four locations

between lagged observations. A non-stationary time series, for instance a time series with trends, seasonality or both, can be modelled using ARIMA, with an appropriate order of differencing. ARIMA models, however, are based on autocorrelation instead of the structural view of level, trend and seasonality.

2.3 | Structural time series models

The advantage of using a structural time series model is the ability to relax the formulation to adapt to changes in series levels over time. The evolutionary nature of the components of climatic time series implies a deviation from the assumption of a fixed pattern and behaviour over time. All the components such as cycle, trend, seasonality and irregular may have stochastic features and complicate the modelling effort. The assumption that these components evolve randomly over time is the basis for structural time series modelling (Jalles, 2009). Here Y_t is defined as:

$$Y_t = \mu_t \mathcal{C}_t \mathcal{S}_t \varepsilon_t \tag{5}$$

where $\mu_t, \mathcal{C}_t, \mathcal{S}_t$ and ε_t are the global trend, cycle, seasonal and irregular components of the time series. The trend and cycle components can be combined into a single level component, and upon applying logarithms the additive decomposition becomes:

$$y_t = \gamma_t + \tau_t + \xi_t \tag{6}$$

where $y_t = \ln Y_t$, $\gamma_t = \ln(\mu_t \mathcal{C}_t), \tau_t = \ln \mathcal{S}_t$ and $\xi_t = \ln \varepsilon_t$. Each component can be considered and formulated as a stochastic process with random disturbances (Pollock, 2008; Jalles, 2009). Each additive component was z-transformed and modelled by an independent ARIMA process, where $\varsigma_\gamma(z)$, $\varsigma_\tau(z)$ and $\xi(z)$ are the z-transforms of independent white noise processes:

$$y_z = \gamma_z + \tau_z + \xi_z \tag{7}$$

$$y_z = \frac{\phi_\gamma(z)}{\theta_\gamma(z)}\varsigma_\gamma(z) + \frac{\phi_\tau(z)}{\theta_\tau(z)}\varsigma_\tau(z) + \xi(z) \tag{8}$$

The z-transformation ensured that the varying signals associated with the components were effectively dealt with and allowed the design of filters for parameter estimation. The unit- root factor $(1 - z)^p$ for the trend component is a factor of the autoregressive polynomial $\theta_\gamma(z)$ whereas the autoregressive polynomial $\theta_\tau(z)$ contains the factor $(1 + z + \cdots + z^{s-1})$, where s represents the number of periods in a seasonal cycle, which was 24, representing daily seasonality. The sum of ARIMA processes yields an ARIMA process (Pollock, 2008) and was represented in the state-space form.

2.4 | ARIMA(p, d, q) in state-space forms, Kalman filter and smoothing

State-space forms offer great flexibility in extracting time series data features. They are generally used for prediction, smoothing and likelihood evaluation purposes (De Jong and Penzer, 2000). They also offer convenient frameworks for incorporating smoothing functions in a wide range of time series models to improve prediction generally. They are specified by two sets of equations, namely the observation and state equations represented in Equations (9) and (10) respectively:

$$y_t = H_t^{\mathrm{T}} Y_t + \varepsilon_t \tag{9}$$

$$Y_t = Z_t Y_{t-1} + R_t \omega_t \tag{10}$$

where, for a state vector of length k, H_t is a vector of length $k, \varepsilon_t \tilde{\mathrm{N}}\mathrm{ID}\left(0, \sigma^2_{\varepsilon_t}\right)$, Z_t is a $k \times k$ matrix, R_t is a $k \times 1$ matrix and $\omega_t \sim \mathrm{NID}(0, W)$. Considering Equation (1), let $m = \max(p + d, q + 1)$. Then:

$$y_t = \theta_1 y_{t-1} + \ldots + \theta_m y_{t-m} + a_t - \phi_1 a_{t-1} - \cdots - \phi_{m-1} a_{t-m+1} \tag{11}$$

Also, for $j < m$ and $\phi_0 = -1$, let:

$$\eta_t^{(m)} = \theta_m y_{t-1} - \phi_{m-1} a_t \tag{12}$$

$$\eta_t^{(j)} = \theta_j y_{t-1} + \eta_{t-1}^{(j+1)} - \phi_{j-1} a_t \tag{13}$$

Then, the state vector $Y_t = \left(\eta_t^{(1)}, \ldots, \eta_t^{(m)}\right)^{\mathrm{T}}$ satisfies

$$Y_t = \begin{bmatrix} \theta_1 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \theta_{m-1} & 0 & \cdots & 1 \\ \theta_m & 0 & \cdots & 0 \end{bmatrix} Y_{t-1} + \begin{bmatrix} 1 \\ -\phi_1 \\ \vdots \\ -\phi_{m-1} \end{bmatrix} \omega_t \tag{14}$$

This form of state-space representation of an ARIMA (p, d, q) model has been found to have both computational and conceptual advantages (Harvey, 1989; Hamilton, 1994). The formulation guarantees that the ARIMA models are responsive to the application of the Kalman filter and smoothing for the estimation of the model parameters and the extraction of unobserved components.

The estimation and updating of the model parameters constitute the Kalman filter and these filters were obtained by considering a Gaussian process with the initial state $Y_t \sim N(\eta_t, V_t)$, with mean and variance at time $t + 1$:

$$\eta_{t+1|t} = Z_{t+1}\eta_t \qquad V_{t+1|t} = Z_{t+1} V_t Z_{t+1}^{\mathrm{T}} + R_{t+1} W_{t+1} \tag{15}$$

The mean and variance of the joint distribution of $\left(Y_{t+1}^{\mathrm{T}}, y_{t+1}^{\mathrm{T}}\right)^{\mathrm{T}}$ are given respectively as:

$$\begin{pmatrix} \eta_{t+1|t} \\ H_{t+1}\eta_{t+1|t} \end{pmatrix} \quad \begin{pmatrix} V_{t+1|t} & V_{t+1|t} H_{t+1}^{\mathrm{T}} \\ H_{t+1} V_{t+1|t} & H_{t+1} V_{t+1|t} H_{t+1}^{\mathrm{T}} + \sigma^2_{\varepsilon_{t+1}} I \end{pmatrix} \tag{16}$$

The mean and variance of the conditional distribution of $Y_{t+1} \mid y_{t+1}$ are obtained as:

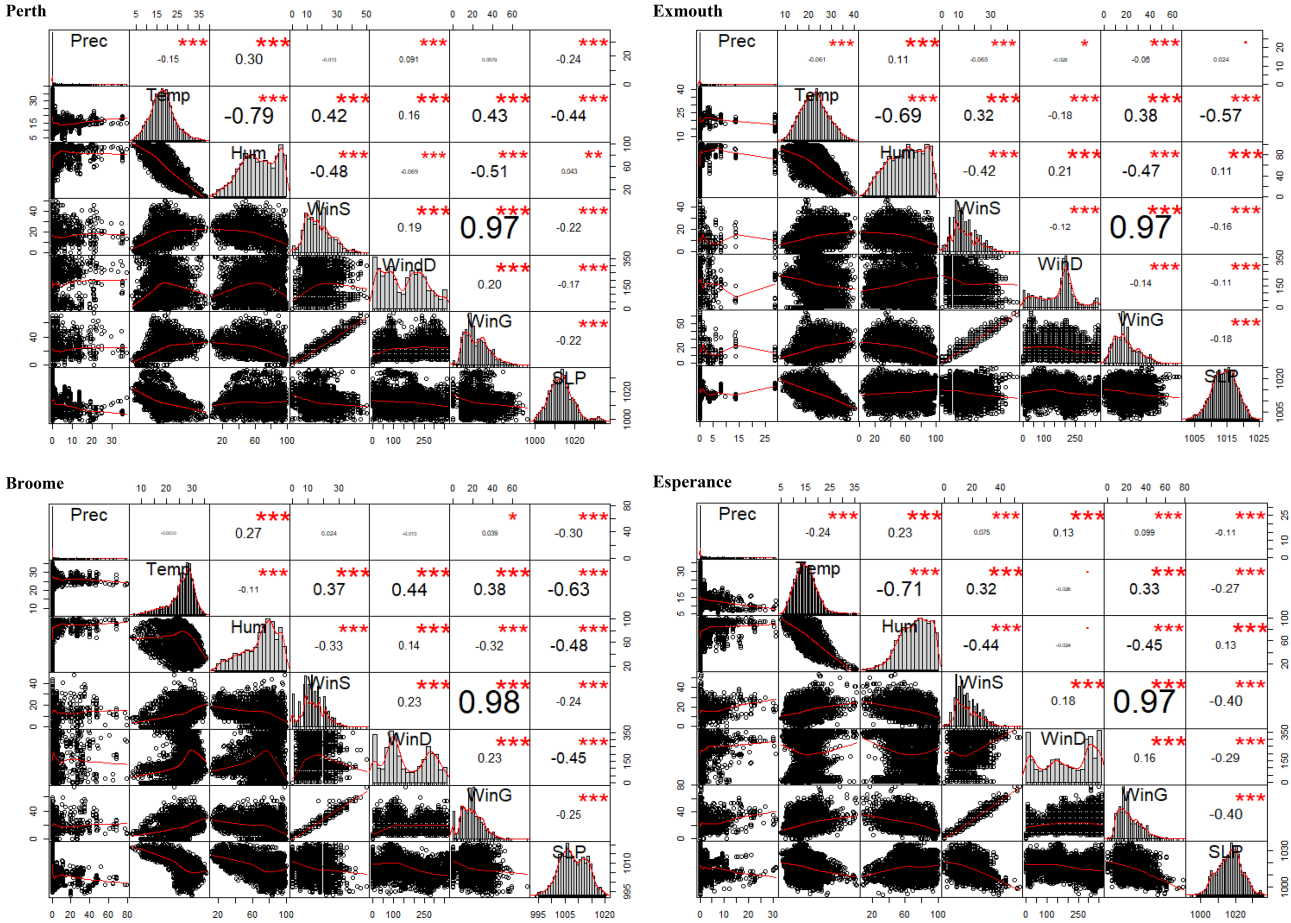

FIGURE 3 Distributional characteristics, paired scatter plots and Pearson correlation between study variables

$$\eta_{t+1} = \eta_{t+1|t} + V_{t+1|t}H_{t+1}^{\mathrm{T}}\left(H_{t+1}V_{t+1|t}H_{t+1}^{\mathrm{T}} + \sigma_{\varepsilon_{t+1}}^2 I\right)^{-1}\left(y_{t+1} - H_{t+1}\eta_{t+1|t}\right) \tag{17}$$

$$V_{t+1} = V_{t+1|t} - V_{t+1|t}H_{t+1}^{\mathrm{T}}\left(H_{t+1}V_{t+1|t}H_{t+1}^{\mathrm{T}} + \sigma_{\varepsilon_{t+1}}^2 I\right)^{-1}H_{t+1}V_{t+1|t} \tag{18}$$

The prediction and the updating of estimates by the Kalman filter can therefore be summarized as follows:

$$\eta_{t+1|t} = Z_{t+1}\eta_t \quad \text{(state prediction)} \tag{19}$$

$$V_{t+1|t} = Z_{t+1}V_t Z_{t+1}^{\mathrm{T}} + R_{t+1}W_{t+1} \quad \text{(prediction dispersion)} \tag{20}$$

$$e_{t+1} = y_{t+1} - H_{t+1}\eta_{t+1|t} \quad \text{(prediction error)} \tag{21}$$

TABLE 2 Multiple linear regression model formulation

Response	Predictors
Temperature	Precipitation, humidity, wind speed, wind direction (sine and cosine transformed), sea level pressure
Humidity	Precipitation, temperature, wind speed, wind direction (sine and cosine transformed), sea level pressure
Wind speed	Precipitation, temperature, humidity, sea level pressure

$$\Sigma_{t+1} = H_{t+1}V_{t+1|t}H_{t+1}^{\mathrm{T}} + \sigma_{\varepsilon_{t+1}}^2 I \quad \text{(error dispersion)} \tag{22}$$

$$K_t = V_{t+1|t}H_{t+1}^{\mathrm{T}}\Sigma_t^{-1} \quad \text{(Kalman gain)} \tag{23}$$

$$\eta_{t+1} = \eta_{t+1|t} + V_{t+1|t}H_{t+1}^{\mathrm{T}}\Sigma_t^{-1}e_{t+1} \quad \text{(state estimate)} \tag{24}$$

$$V_{t+1} = V_{t+1|t} - V_{t+1|t} H_{t+1}^{\mathrm{T}} \Sigma_t^{-1} H_{t+1} V_{t+1|t} \quad \text{(estimate dispersion)} \tag{25}$$

The Kalman filter operates recursively where current best estimates are updated when new observations are available. These models were implemented using the "imputeTS" package, version 2.7 (Moritz and Bartz-Beielstein, 2017), making use of the options "StructTS" and "auto.arima" of the function "na.kalman" in the R software version 3.5.1 (R Core Team, 2013).

2.5 | Multiple regression modelling

The multiple regression models in this study were formulated with both continuous and circular predictors. Precipitation, temperature, humidity, wind gust and sea level pressure were treated as continuous variables. Following SenGupta and Ugwuowo (2006), wind direction as a circular variable was trigonometrically transformed by:

$$\cos\left(\pi \frac{\text{direction}}{180}\right), \sin\left(\pi \frac{\text{direction}}{180}\right)$$

The resultant variables representing wind speed together with the continuous variables were used in the multiple regression modelling. The estimate of a missing observation at time t was imputed using the model:

$$\hat{y}_t = \hat{\beta}_{t_0} + \sum_{i=1}^{p} \hat{\beta}_{t_i} X_{t_i} \tag{26}$$

where X represents the vector of p predictors with parameters β.

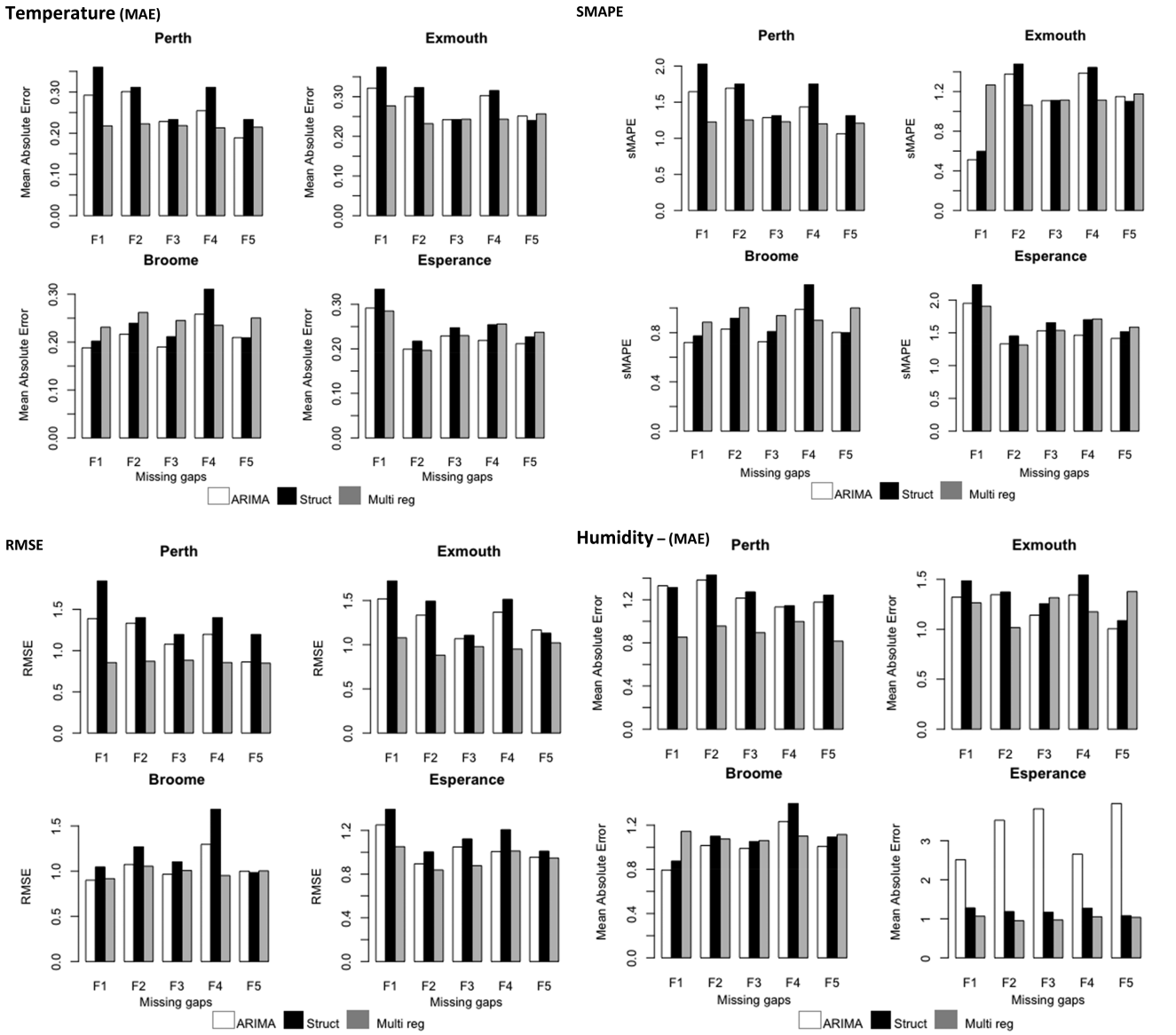

FIGURE 4 The mean absolute error, root mean square error and symmetric mean absolute error values for the performance evaluation of imputing the five missing folds of temperature, humidity and wind speed from the different estimation techniques for the four study locations

To overcome the possible violation of the model assumption of non-autocorrelated errors, heteroscedasticity and autocorrelation robust standard errors were used (Newey and West, 1987). For imputation purposes, independent variables that are highly correlated with a dependent variable with missing observations can be modelled to obtain highly plausible imputations (van Buuren and Groothuis-Oudshoorn, 2011). Figure 3 depicts the distributional characteristics and the nature and strength of the relationship between the study variables. The distributional characteristics, the nature and strength of the relationship differed with respect to location. The strength of the relationships could inform how important a predictor was in the regression imputation models. For instance, from Figure 3, except for Broome, high correlation values were observed between temperature and humidity and each variable would serve as a good predictor for imputing the other. The nature of the relationship between variables presented in Figure 3 especially between the resultant variables (sine and cosine) of transformation does not look linear. However, the linear regression modelling technique according to Berman (1998) has a non-parametric interpretation as the average linear trend across all pairs of observations and the relationships between variables do not have to look linear and could be applied to any monotonic trend. The multiple linear regression models were formulated with the response and predictor variables depicted in Table 2.

3 | MODEL PERFORMANCE EVALUATION

Using the out-of-sample splits in the five-fold cross-validation, the estimation accuracy of the models in the

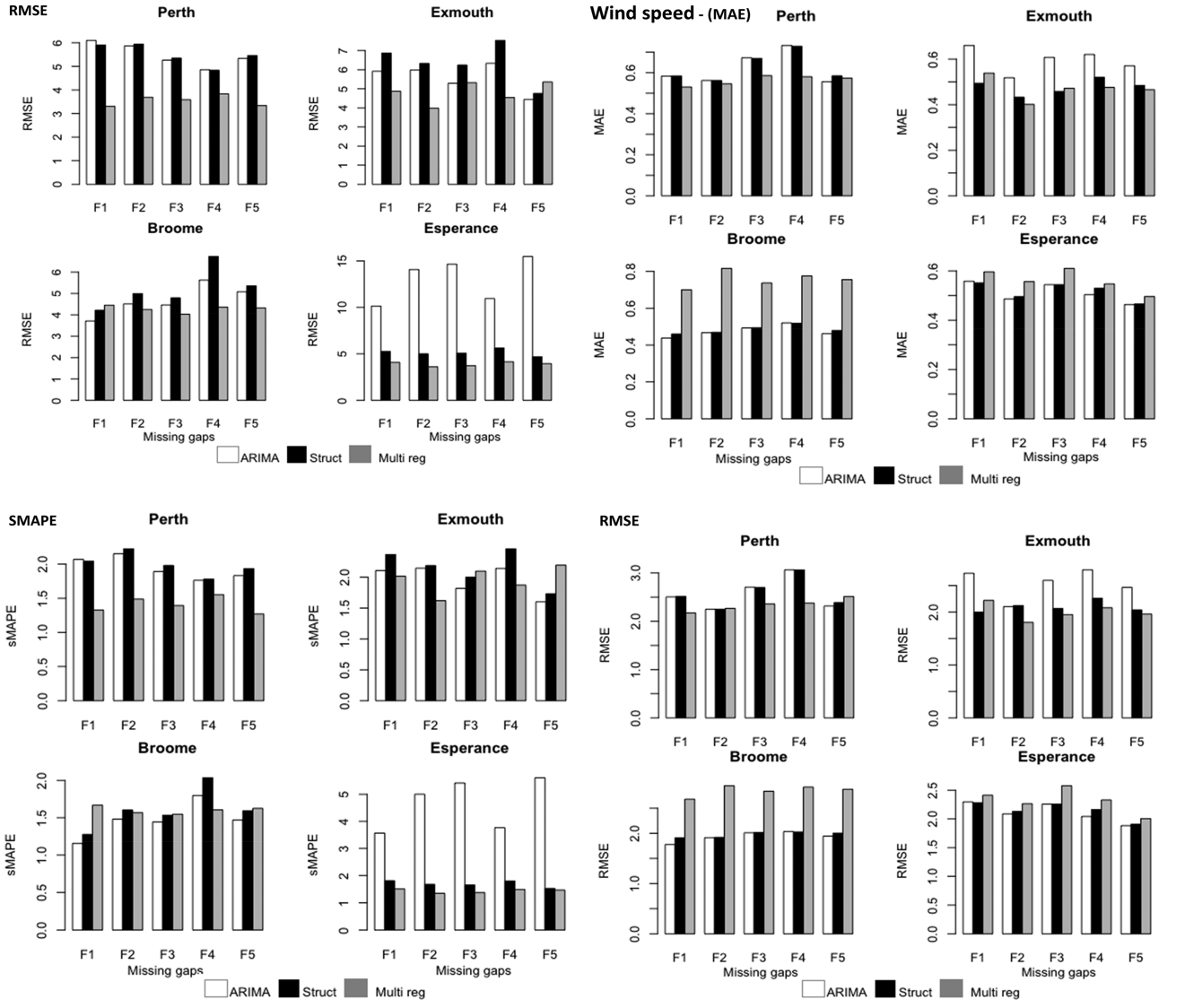

FIGURE 4 (Continued)

imputations was assessed by means of three model performance metrics: the MAE, RMSE and the symmetric mean absolute percentage error (SMAPE). Equations (27)–(29) respectively represent the metrics:

$$\mathrm{MAE} = \frac{\sum_{t=1}^{T} |\hat{y}_t - y_t|}{T} \tag{27}$$

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{t=1}^{T} (\hat{y}_t - y_t)^2}{T}} \tag{28}$$

$$\mathrm{SMAPE} = \frac{\sum_{t=1}^{T} |(\hat{y}_t - y_t)/y|}{T} \times 100 \tag{29}$$

where $\bar{y} = \sum_{t=1}^{T} y_t / T$, $\hat{y}_t$ and y_t are the imputed and actual values for time t and T is the length of the missing data in the time series. The MAE and RMSE are useful performance indicators (McCandless *et al*., 2011; Kanda *et al*., 2018; Lara-Estrada *et al*., 2018). Both the MAE and RMSE range from 0 to $+\infty$, and lower values indicate high levels of agreement between observed and estimated values. The MAE and RMSE values have the same units as the variables measured. SMAPE is a variation of mean absolute percentage error (MAPE) proposed by Makridakis (1993) where, instead of dividing the error by the observed value, the observed average is used instead. This ensures that symmetry is achieved and thus is applicable to variables with meaningful zero values, for instance a 0°C temperature value. Smaller percentage values indicate high levels of agreement between observed and imputed values.

4 | RESULTS AND DISCUSSION

The estimation accuracy of the methods was assessed based on the performance indicators described in the preceding section. Generally, the performance indicators agreed in the selection of the best imputation method in the five-fold gaps of missing observations (see Figure 4). There were instances, however, where the measures ranked the methods differently, notably the RMSE compared to the other indicators. For example, in assessing the imputations of wind speed in Perth, the methods were ranked differently for fold 2, while both the MAE and SMAPE ranked multiple linear regression as the best and RMSE ranked the structural model with Kalman smoothing as the best. Similarly, for humidity in Broome, RMSE ranked multiple linear regression as the best while the MAE and SMAPE both ranked the ARIMA with Kalman smoothing as the best method in folds 2, 3 and 5 (see Figure 4). In cases where the indicators followed divergent paths in ranking the models, the MAE was used in assessing the models (Legate and McCabe, 1999; Kanda *et al*., 2018). The best methods were determined based on the

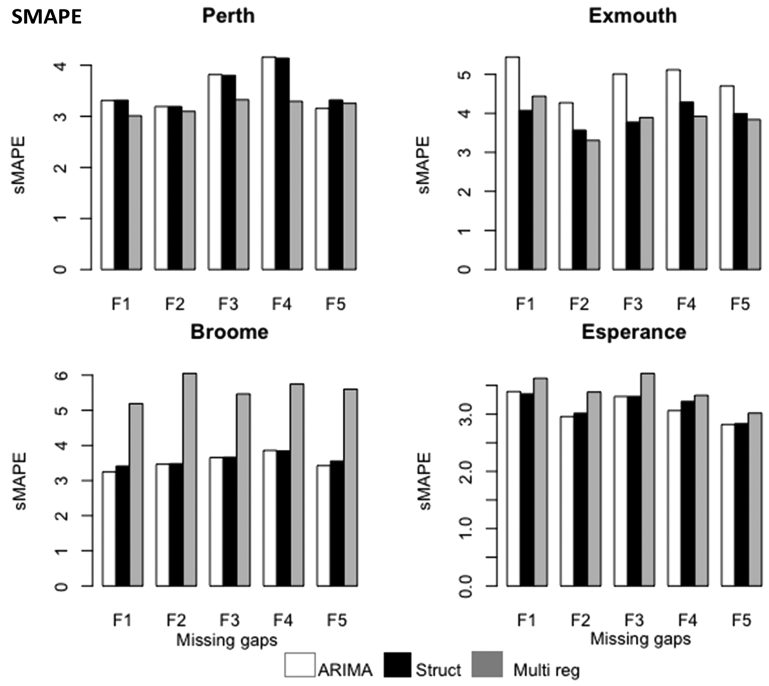

FIGURE 4 (Continued)

pooled averaged values of the performance indicators over the five folds.

Table 3 presents the pooled performance indicators' average values for the five-fold missing observations and the rankings of the methods in imputing missing values for temperature, humidity and wind speed at the four locations. Multiple linear regression was consistently ranked as the best among the methods especially for the more southerly locations Perth and Esperance. The performance of multiple linear regression at these locations could be attributed to the relatively strong relationships between the studied variables compared to the strength

TABLE 3 Model ranking based on the pooled average performance indicators' values across the five-fold gaps of missing observations

Variable	Location	Method	MAE	RMSE	SMAPE
Temperature	Esperance	ARIMA	**0.23**	1.0308	**1.5385**
		Structural	0.2559	1.1469	1.7113
		Multiple	0.2409	**0.9448**	1.611
	Perth	ARIMA	0.2529	1.1713	1.4236
		Structural	0.299	1.407	1.6309
		Multiple	**0.2173**	**0.8616**	**1.2231**
	Exmouth	ARIMA	0.2836	1.2906	**1.1073**
		Structural	0.299	1.3925	1.1463
		Multiple	**0.2504**	**0.9814**	1.1471
	Broome	ARIMA	**0.2124**	1.0465	**0.8121**
		Structural	0.2344	1.2166	0.8973
		Multiple	0.2446	**0.991**	0.945
Humidity	Esperance	ARIMA	3.2916	13.05	4.6726
		Structural	1.1948	5.1328	1.6961
		Multiple	**1.0137**	**3.8977**	**1.439**
	Perth	ARIMA	1.2475	5.4855	1.9418
		Structural	1.2802	5.5	1.9927
		Multiple	**0.9033**	**3.5517**	**1.4061**
	Exmouth	ARIMA	1.2315	5.5933	1.9621
		Structural	1.3486	6.3466	2.1485
		Multiple	**1.23**	**4.815**	**1.9596**
	Broome	ARIMA	**1.0078**	4.6773	**1.4698**
		Structural	1.1024	5.2174	1.6078
		Multiple	1.0988	**4.2813**	1.6025
Wind speed	Esperance	ARIMA	**0.5112**	**2.1135**	**3.1073**
		Structural	0.5173	2.1484	3.1449
		Multiple	0.5613	2.3183	3.4122
	Perth	ARIMA	0.6217	2.5687	3.5287
		Structural	0.6262	2.5833	3.5544
		Multiple	**0.5631**	**2.3364**	**3.1963**
	Exmouth	ARIMA	0.5955	2.5379	4.9074
		Structural	0.4779	2.0965	3.9378
		Multiple	**0.4707**	**2.0024**	**3.8787**
	Broome	ARIMA	**0.4763**	**1.9354**	**3.5308**
		Structural	0.4845	1.9752	3.5916
		Multiple	0.7568	2.8519	5.6109

Note: The most appropriate models as ranked by the performance indicators for the study variables with respect to location are in bold and the underlined text were ranked second.

of the relationships in the northern parts of WA (see Figure 3). However, in Broome (in the north of WA) multiple linear regression was ranked as the worst model for all the variables except for humidity, based on the MAE. From Figure 3, the Pearson correlation coefficients between the variables suggest a very weak relationship between the variables in Broome and possibly contributed to the poor performance of the multiple regression model (Petris, 2010; van Buuren and Groothuis-Oudshoorn, 2011). Broome has a tropical climate defined by dry and wet seasons, which is a complete departure from the typical four seasons, namely summer, winter, autumn and spring, in the other locations. For instance, the Pearson correlation between temperature and humidity was $r = -0.111$, significantly lower than the strength of correlation in the other locations (see Figure 3).

The locations appeared to influence the choice of the preferred model. Kashani and Dinpasho (2012) concluded that climate data in different land topologies were significantly different. The geomorphological and vegetation features are significantly different between these locations (found in different climate zones) and will interact differently in the mechanisms leading to the formations of these climatic variables (Kanda *et al.*, 2018). The nature of the relationships between the study variables in Figure 3 also suggests that the data structures in the locations are different. The choice of model was guided by the data characteristics which were apparently influenced by the locations. In the models' choice, the ARIMA with Kalman smoothing performed best in Broome which is characterized by hot semi-arid desert conditions (Figure 1). Similarly, multiple linear regression performed best in Perth in the hot summer Mediterranean zone for all the variables studied. The density plots (see Figures 5–8) of temperature and wind speed suggest that the series were to some extent similar for the locations in the southern part of WA (Esperance and Perth) and similarly for those in the north (Exmouth and Broome).

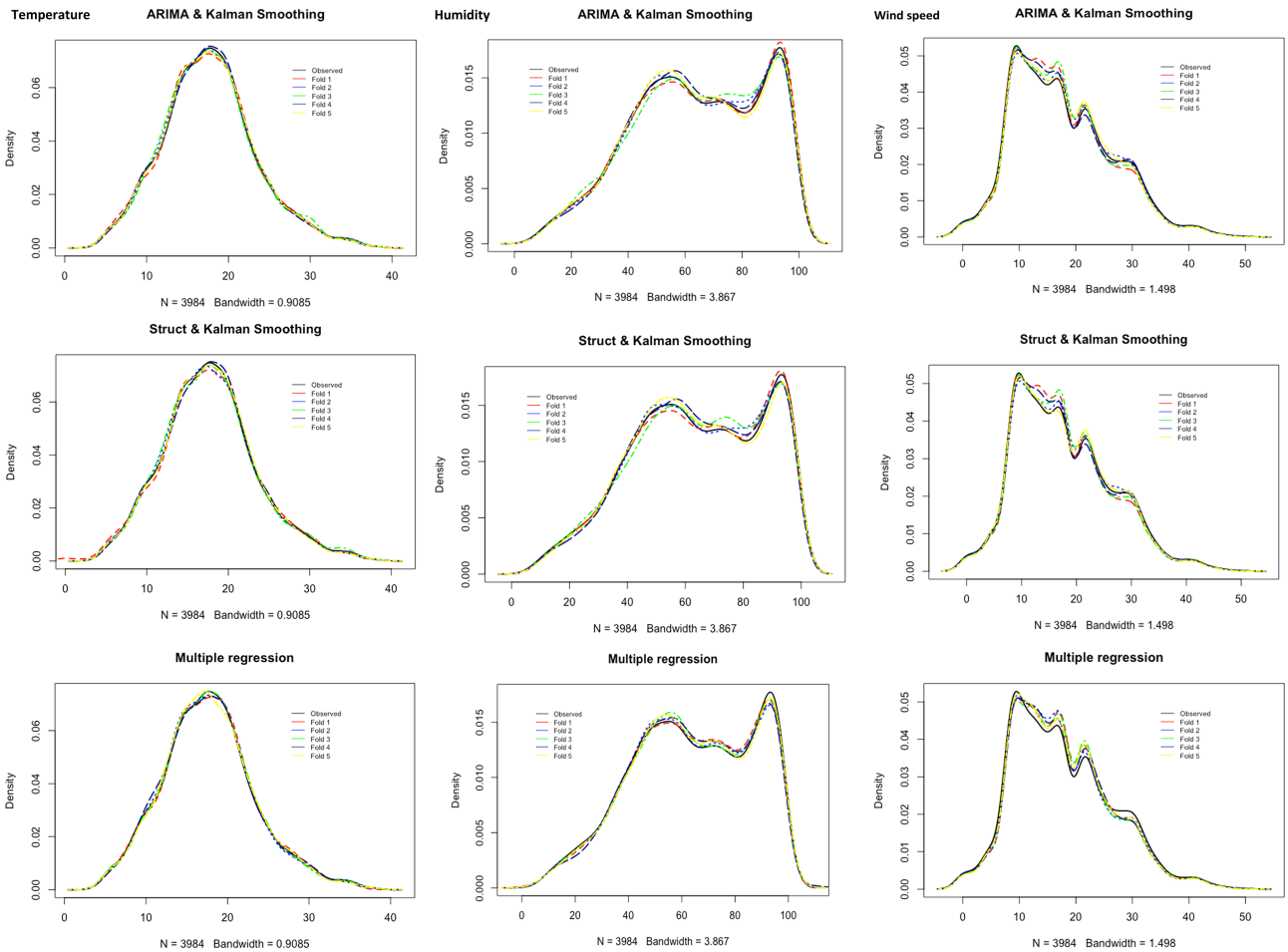

FIGURE 5 Density plots comparing the distribution of the observed and the five missing folds of imputed values for temperature, humidity and wind speed in Perth

From Table 3, the average MAE associated with the imputed temperature missing values by the methods was 0.25°C with a magnitude of error less than 0.30°C for all the four locations. The difference in the MAE values between the top competing models was 0.01, 0.04,0.03 and 0.02°C for Esperance, Perth, Exmouth and Broome, respectively. For humidity, the average MAE was 1.320% with a magnitude of error of less than 3.400% for all locations. The differences in the MAE values between the top two models were 0.180%, 0.340%, 0.002% and 0.090% for Esperance, Perth, Exmouth and Broome, respectively. In the case of wind speed, the average MAE value association with the imputations was 0.560 $km \cdot h^{-1}$ with a magnitude of error less than 0.900 $km \cdot h^{-1}$ for all locations. For the two best performing models, the differences between the MAE values were 0.006, 0.060, 0.007 and 0.008 $km \cdot h^{-1}$ for Esperance, Perth, Exmouth and Broome respectively. The relatively low magnitude of error values for the variables suggested that the methods were closely competing and imputed highly plausible values.

To check the validity and plausibility of the imputations performed by the methods further, the density plots of the imputations of the five-fold gaps of missing observations were overlaid on that of the observed. Figures 5–8 depict the overlaid density plots for the temperature, humidity and wind speed with respect to location. The preservation of the distribution of the variable with missing values to enable the derivation of reliable sample estimators is a key objective of data imputation (Brick and Kalton, 1996). The density plots showed a close resemblance between the distribution of the observed and the imputed values for the study variables in the four locations. Figures 5–8 suggest that the imputed values were highly plausible to allow any analyses to be conducted as if the data were complete. The resemblance of the imputed values in the five-fold gaps of missing observations to the observed data was also assessed using the correlation coefficient (R) (see Table 4). The coefficients of determination scores observed by the methods were $0.946 \leq R \leq 0.988$,

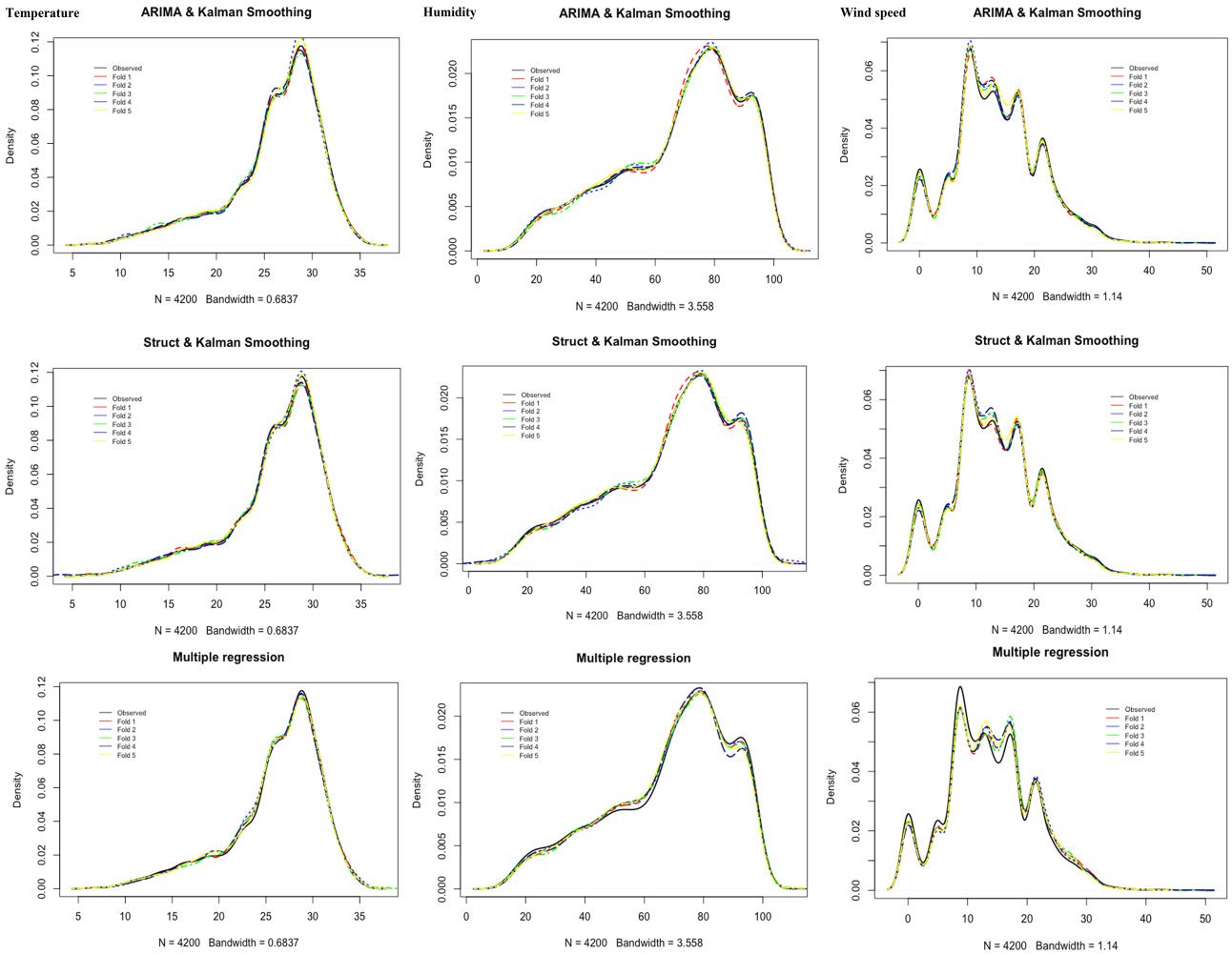

FIGURE 6 Density plots comparing the distribution of the observed and the five missing folds of imputed values for temperature, humidity and wind speed in Broome

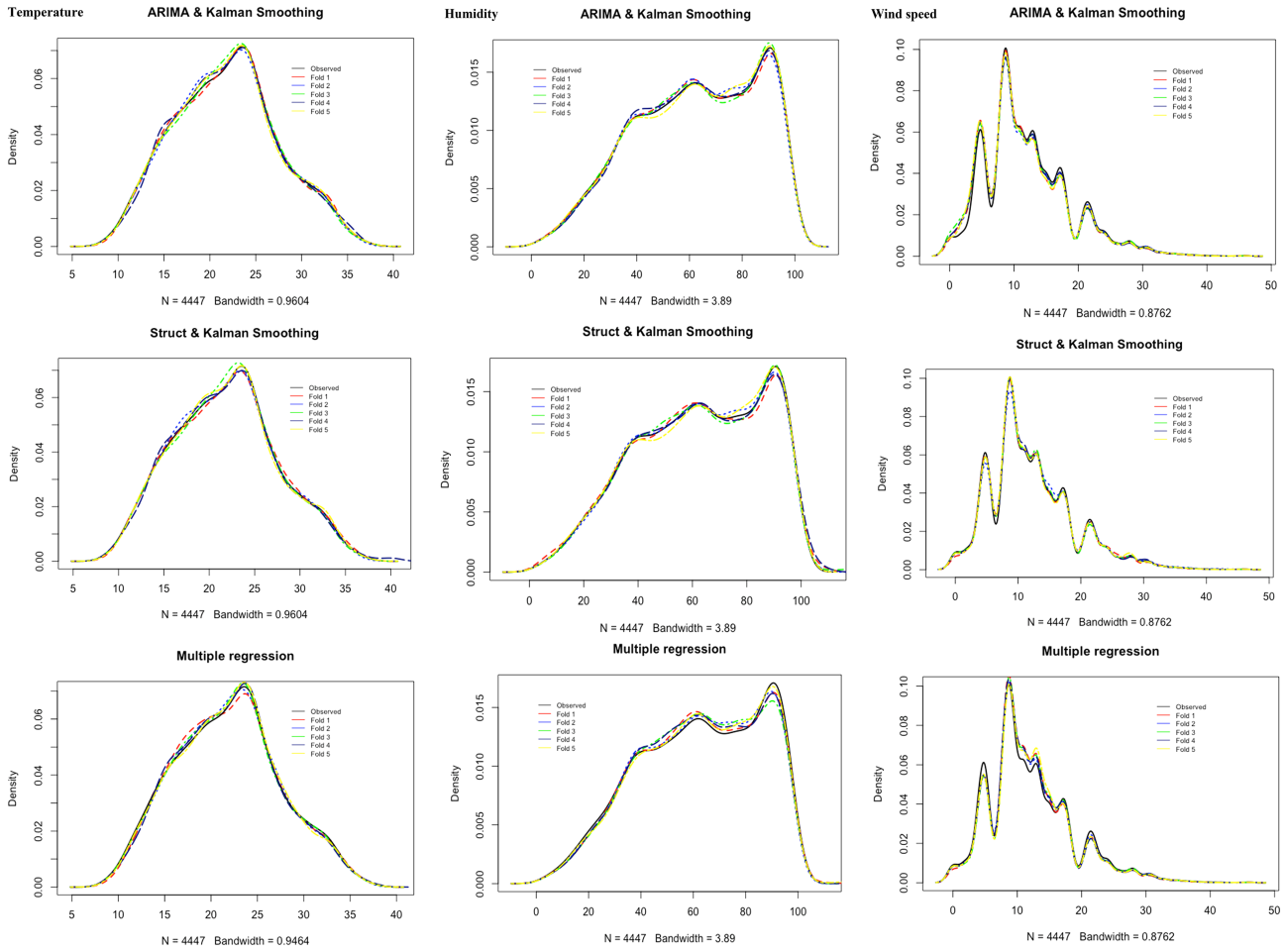

FIGURE 7 Density plots comparing the distribution of the observed and the five missing folds of imputed values for temperature, humidity and wind speed in Exmouth

$0.747 \leq R \leq 0.989$ and $0.957 \leq R \leq 0.971$ for temperature, humidity and wind speed respectively for the locations. Except, for humidity in Esperance where the ARIMA with Kalman smoothing seemed more biased, recording low scores of $0.747 \leq R \leq 0.872$, the models generally recorded high R values suggesting that the imputed values in the five-fold gaps of missing observations by the models were very similar to the observed values.

The multiple linear regression model was the best model in most cases, followed by the ARIMA with Kalman smoothing. Multiple linear regression models of highly correlated response and predictor variables have been found to impute highly plausible values (van Buuren and Groothuis-Oudshoorn, 2011; Kashani and Dinpasho, 2012; Kanda *et al.*, 2018). The comparatively poor performance of the structural time series models to the ARIMA models could be due to the fact that the underlying structures of the series studied were simple and had relatively short cycles. The lengths of the series studied were relatively short and did not exhibit any complexity in structure. For instance, the series only cut across two and three seasons and in effect features such as trend, seasonality and cyclic behaviours were obscured. This reduced the level of complexity of the time series and to some extent undermined the potential of the structural time series models. The structural time series models are known to be efficient in modelling series with complex structure, where prior analysis of the structure underlying the generating system is carried out before the model fit (Jalles, 2009). The structural time series model explicitly modelled the trend, seasonal, error terms and other relevant components. There were some cases especially with wind speed in Exmouth where the structural model outperformed the ARIMA model. Observing the distributional characteristics of the variables in Figures 5–8 suggests that the series for wind speed was more erratic compared to temperature and humidity. The structural models, although ranked as the worst performing models in most cases, still performed

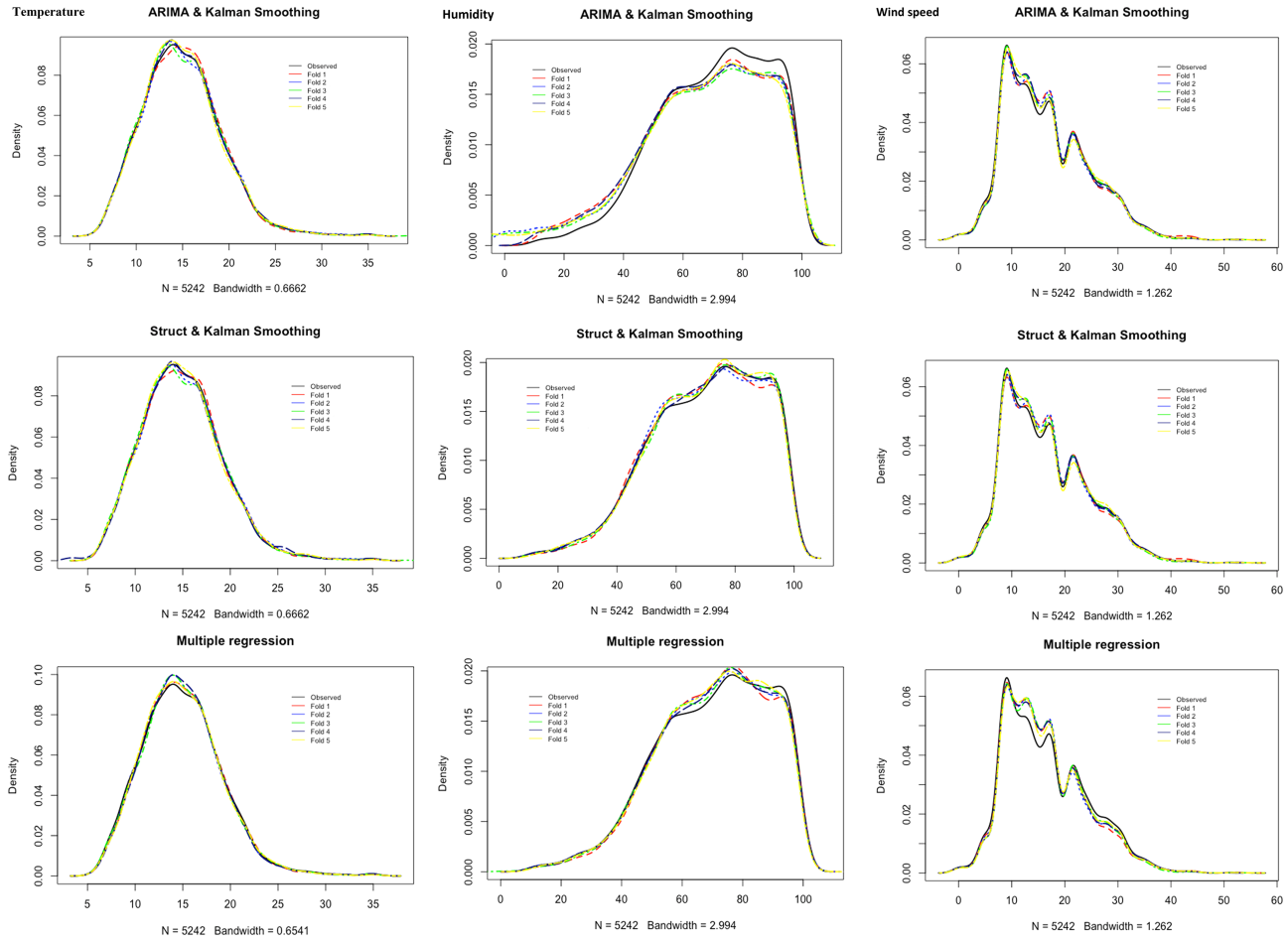

FIGURE 8 Density plots comparing the distribution of the observed and the five missing folds of imputed values for temperature, humidity and wind speed in Esperance

satisfactorily in modelling and imputing the missing observations in relatively short climate time series datasets.

4.1 | Limitations and opportunities of the modelling techniques

Missing data are common in instrumentation measurements and missingness can be persistent (Simolo *et al.*, 2010). The choice of imputation technique, all things being equal, should allow for an easy and efficient modelling technique. In this study, the multiple linear regression model was the least complex model and in most cases performed best. However, in the absence of predictors or if missing data are persistent in predictors, the method cannot be used. Data scarcity is common especially in developing countries and access to datasets for important predictor variables may not be feasible. Also, the modelling assumptions need to be considered, and if they are not satisfied the use of multiple linear regression modelling may not be applicable. The use of heteroscedasticity and autocorrelation may have worked well because of the lengths of the series studied; for relatively long series it could easily result in model misspecification leading to less accurate predictions. Alternatively, the univariate time series models by the state-space method, despite the level of complexity, are time dependent and are independent of other variables. The techniques competed closely with multiple linear regression even for very short cycle time series datasets. The plausibility of the imputed values obtained from these techniques makes them ideal for any situation. The advances in statistical software and packages would enable easy implementation of these techniques.

TABLE 4 The correlation coefficient between the imputed values and the observed values with respect to the estimation techniques for the five missing folds

Variables	Method	Location	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Temperature	ARIMA and Kalman smoothing	Esperance	0.956	0.978	0.97	0.972	0.974
		Perth	0.971	0.974	0.983	0.978	0.989
		Exmouth	0.965	0.972	0.982	0.971	0.979
		Broome	0.984	0.977	0.981	0.966	0.98
	Structural and Kalman smoothing	Esperance	0.946	0.973	0.965	0.96	0.972
		Perth	0.95	0.97	0.979	0.977	0.974
		Exmouth	0.955	0.966	0.981	0.966	0.981
		Broome	0.978	0.967	0.976	0.946	0.98
	Multiple regression	Esperance	0.969	0.98	0.978	0.971	0.975
		Perth	0.989	0.988	0.988	0.989	0.989
		Exmouth	0.982	0.988	0.985	0.986	0.984
		Broome	0.983	0.977	0.979	0.981	0.985
Humidity	ARIMA and Kalman smoothing	Esperance	0.873	0.783	0.781	0.851	0.747
		Perth	0.964	0.966	0.973	0.977	0.972
		Exmouth	0.967	0.966	0.974	0.962	0.982
		Broome	0.984	0.977	0.977	0.963	0.97
	Structural and Kalman smoothing	Esperance	0.959	0.963	0.961	0.953	0.967
		Perth	0.966	0.965	0.972	0.977	0.97
		Exmouth	0.957	0.962	0.964	0.948	0.979
		Broome	0.978	0.972	0.974	0.949	0.967
	Multiple regression	Esperance	0.975	0.981	0.979	0.975	0.977
		Perth	0.989	0.987	0.987	0.985	0.989
		Exmouth	0.978	0.985	0.973	0.981	0.973
		Broome	0.977	0.979	0.981	0.978	0.979
Wind speed	ARIMA and Kalman smoothing	Esperance	0.956	0.963	0.957	0.964	0.97
		Perth	0.958	0.966	0.951	0.938	0.964
		Exmouth	0.92	0.953	0.929	0.915	0.935
		Broome	0.971	0.966	0.925	0.925	0.931
	Structural and Kalman smoothing	Esperance	0.916	0.925	0.916	0.924	0.941
		Perth	0.918	0.933	0.904	0.88	0.925
		Exmouth	0.912	0.904	0.906	0.887	0.908
		Broome	0.933	0.933	0.962	0.962	0.963
	Multiple regression	Esperance	0.952	0.957	0.944	0.954	0.966
		Perth	0.969	0.966	0.963	0.962	0.958
		Exmouth	0.944	0.963	0.957	0.951	0.957
		Broome	0.937	0.923	0.929	0.923	0.924

5 | CONCLUSION

The methods studied have demonstrated their suitability in imputing missing data in high-resolution temperature, humidity and wind speed data. However, the study only used sub-samples of relatively short time series and this could have contributed to the general performance of the univariate time series modelling approaches. It is recommended that longer climate time series datasets with varying patterns of complexity are studied to assess the

techniques further under several varying scenarios of missing data. The assumption that data were missing at random may not be applicable to all causes of missing data, for instance missingness caused by routine maintenance or calibration. To ensure that the missing at random assumption holds, it is recommended that such activities are carried out in a randomized fashion.

ACKNOWLEDGEMENT

This study was funded and supported by the Government of Western Australia Department of Primary Industries and Regional Development (DPIRD). The authors are grateful to the Australian Bureau of Meteorology for providing the data for the study. The authors wish to thank Dr Ainslie Denham, Dr Timothy Green and Mr Mark Pagano for their input during the internal review process by DPIRD. Thanks are also due to the two anonymous reviewers for their constructive criticisms and the useful suggestions made.

ORCID

E Afrifa-Yamoah https://orcid.org/0000-0003-1741-9249

REFERENCES

Berman, M. (1998) A theorem of Jacobi and its generalization. *Biometrika*, 75(4), 779–783.

Bianchi, M., Boyle, M. and Hollingsworth, D. (1999) A comparison of methods of trend estimation. *Applied Economics Letters*, 6, 103–109.

Brick, J.M. and Kalton, G. (1996) Handling missing data in survey research. *Statistical Methods in Medical Research*, 5(3), 215–238.

Coble, J., Ramuhalli, P., Meyer, R., Hashemian, H.M., Shumaker, B. and Cummins, D. (2012) 'Calibration monitoring for sensor calibration interval extension: identifying technical gaps'. *Future of Instrumentation International Workshop (FIIW) Proceedings*. Gatlinburg, TN: IEEE.

de Jong, P. and Penzer, J. (2000) The ARIMA model in state-space form. Department of Statistics, London School of Economics. *Research Report*, 40, 1–10.

de Jong, R., van Buuren, S. and Spiess, M. (2016) Multiple imputation of predictor variables using generalized additive models. *Communications in Statistics – Simulation and Computation*, 45, 968–985.

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977) Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39, 1–38.

Firat, M., Dikbas, F., Cem Koc, A. and Gungor, M. (2012) Analysis of temperature series: estimation of missing data and homogeneity test. *Meteorological Applications*, 19, 397–406.

Harvey, A.C. (1989) *Forecasting, structural time series models and the Kalman filter*. Cambridge, UK: Cambridge University Press.

Harvey, A.C. and Peters, S. (1990) Estimation procedures for structural time series models. *Journal of Forecasting*, 9, 89–108.

Hamilton, J.D. (1994) *Time Series Analysis*. Princeton, NJ: Princeton University Press.

Ho, S.L. and Xie, M. (1998) The use of ARIMA models for reliability forecasting and analysis. *Computers and Industrial Engineering*, 35, 213–216.

Hyndman, R.J. and Fan, S. (2015) Monash Electricity Forecasting Model. Report for Australian Energy Market Operator (AEMO). Available at: http://robjhyndman.com/papers/MEFMR1.pdf. [].

Jalles, J.T. (2009) *Structural time series models and Kalman filter: a concise review*. FEUNL Work Paper 541. http://dx.doi.org/10.2139/ssrn.1496864

Kanda, N., Negi, H.S., Rishi, M.S. and Shekhar, M.S. (2018) Performance of various techniques in estimating missing climatological data over snowbound mountainous areas of Karakoram Himalaya. *Meteorological Applications*, 25, 337–349.

Kashani, M.H. and Dinpasho, Y. (2012) Evaluation of efficiency of different estimation methods for missing climatological data. *Stochastic Environmental Research and Risk Assessment*, 26, 59–71.

Lara-Estrada, L., Rasche, L., Sucar, E. and Schneider, U.A. (2018) Inferring missing climate data for agricultural planning using Bayesian network. *Land*, 7(4), 1–13.

Legates, D.R. and McCabe, G.J. (1999) Evaluating the use of 'goodness of fit' measures in hydrologic and hydroclimatic model validation. *Water Resources Research*, 35(1), 233–241.

Makridakis, S. (1993) Accuracy measures: Theoretical and practical concerns. *International Journal of Forecasting*, 17(4), 527–529.

McCandless, T.C., Haupt, S.E. and Young, G.S. (2011) The effects of imputing missing data on ensemble temperature forecasts. *Journal of Computers*, 6, 162–171.

Moritz, S., Sardá, A., Bartz-Beielstein, T., Zaefferer, M. and Stork, J. (2015) Comparison of different methods for univariate time series imputation in R. *ArXiv e-print*, Oct. 2015.

Moritz, S. and Bartz-Beielstein, T. (2017) imputeTS: Time series missing value imputation in R. *R Journal*, 9(1), 207–218.

Newey, W.K. and West, K.D. (1987) A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55(3), 703–708.

Pang, X., Shaw, M.D., Lewis, A.C., Capenter, L.J. and Batchellier, T. (2017) Electrochemical ozone sensors: a miniaturised alternative for ozone measurements in laboratory experiments and air-quality monitoring. *Sensors and Actuators B: Chemical*, 240, 829–837.

Petris, G. (2010) An R package for dynamic linear models. *Journal of Statistical Software*, 36(12), 1–16.

Pincetl, S., Graham, R., Murphy, S. and Sivaraman, D. (2015) Analysis of high-resolution utility data for understanding energy use in urban systems: the case of Los Angeles, California: electricity use in Los Angeles. *Journal of Industrial Ecology*, 20(1), 166–178.

Pollock, D.S.G. (2008) Statistical Signal Extraction and Filtering: Structural Time Series Models. Available at: https://www.le.ac.uk/users/dsgp1/ERCSTUFF/ercimstruct.pdf [Accessed 3rd August 2018].

R Core Team. (2013) *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rubin, D.B. (1976) Inference and missing data. *Biometrika*, 63(3), 581–592.

Ryan, K.L., Hall, N.G., Lai, E.K., Smallwood, C.B., Taylor, S.M. and Wise, B.S. (2015) State-wide survey of boat-based recreational fishing in Western Australia 2013/14. Fisheries Research Report 286: Department of Fisheries, Western Australia, 168pp

Schneider, T. (2001) Analysis of incomplete climate data: estimation of mean values and covariance matrices and imputation of missing values. *Journal of Climate*, 14, 853–871.

SenGupta, A. and Ugwuowo, F.I. (2006) Asymmetric circular-linear multivariate regression models with applications to environmental data. *Environmental and Ecological Statistics*, 13, 299–309.

Simolo, C., Brunetti, M., Maugeri, M. and Nanni, T. (2010) Improving estimation of missing values in daily precipitation series by a probability density function-preserving approach. *International Journal of Climatology*, 30, 1564–1576.

Steffe, A.S., Taylor, S.M., Blight, S.J., Ryan, K.L., Desfossess, C., Tate, A., Smallwood, C.B., Lai, E.K., Trinnie, F.I. and Wise, B.S. (2017) Framework for Integration of Data from Remotely Operated Cameras into Recreational Fishery Assessments in Western Australia. *Fisheries Research Report* No. 286, Department of Primary Industries and Regional Development (DPIRD), WA.

Smaragdis, P., Raj, B. and Shashanka, M. (2009) 'Missing data imputation for spectral audio signals'. In *Proc. MLSP*, Grenoble, France, Sep. 2009.

van Buuren, S. and Groothuis-Oudshoorn, K. (2011) MICE: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, 45, 1–67.

Xu, C.D., Wang, J.F., Hu, M.G. and Li, Q.X. (2013) Interpolation of missing temperature data at meteorological stations using P-BSHADE*. *Journal of Climate*, 26, 7452–7463.

World Bank. (2012) *Turn Down the Heat: Why a 4°C Warmer World must be Avoided*. Washington, DC: World Bank. Available at: https://openknowledge.worldbank.org/handle/10986/11860 [Accessed 1st August 2018].

Yodah, W.O., Kihoro, J.M., Athiany, K.H.O. and Kibunja, H.W. (2013) Imputation of incomplete non-stationary seasonal time series data. *Mathematical Theory and Modeling*, 3(12), 142–154.

Yozgatligil, C., Aslan, S., Iyigun, C. and Batmaz, I. (2013) Comparison of missing value imputation methods in time series: the case of Turkish meteorological data. *Theoretical and Applied Climatology*, 112, 143–167.

How to cite this article: Afrifa-Yamoah E, Mueller UA, Taylor SM, Fisher AJ. Missing data imputation of high-resolution temporal climate time series data. *Meteorol Appl*. 2020;27:e1873. https://doi.org/10.1002/met.1873